Experimental Evaluation of Stereo-Based Person Tracking in Real Environment

Junji Satake Jun Miura
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan
satake@cs.tut.ac.jp

Abstract

This paper presents an experimental evaluation of person following in real environment. We built a mobile robot system which follows a specific person while avoiding obstacles and other people walking around. Each person is detected and tracked by simple template matching using distance information obtained by stereo. In this paper, we perform experiments of specific person following in real environment where many ordinary people exist, and analyze the results.

1 Introduction

There is an increasing demand for service robots operating in public space like a shopping mall. An example of service task is to follow a person with carrying his/her items. We developed a robot system that can follow a specific user among obstacles and other people. This paper presents an experimental evaluation in real environment.

There have been a lot of works on person detection and tracking using various image features. HOG [1] is currently one of the most widely used features for visual people detection. Moreover, the person detection methods which combined HOG and the distance information acquired using an RGB-D camera such as Kinect sensor are also proposed [2, 3, 4]. Spinello et al. [2] performed an experiment using fixed cameras in the lobby of an university canteen. Munaro et al. [3] showed an example of tracking result using a mobile robot in an exhibition. Kinect sensor, however, cannot be used under sunlight. Ess et al. [5] proposed to integrate various cues such as appearance-based object detection, depth estimation, visual odometry, and ground plane detection using a graphical model for pedestrian detection. Their method exhibits a nice performance for complicated scenes where many pedestrians exist. However, it is still costly to be used for controlling a real robot. Frintrop et al. [6] proposed a visual tracker for mobile platforms, but their experiments were performed in only laboratory environments.

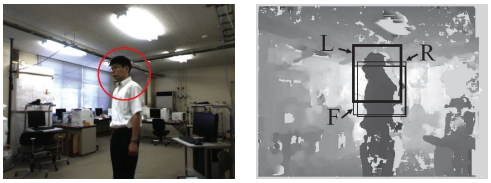
In the previous work [7, 8], we built a mobile robot system with a stereo camera and a laser range finder, and proposed a method of person tracking using distance information obtained by stereo. Although our approach is relatively simple, our robot could robustly follow a specific person while recognizing the target and other persons with occasional occlusions. In this paper, we perform experiments of specific person following in real environment where many ordinary people exist, and analyze the results.

2 Stereo-based person tracking

2.1 Depth template-based person detection

To track persons stably with a moving camera, we use *depth templates* [7], which are the templates for human upper bodies in depth images (see Fig. 1). We made the templates manually from the depth images where the target person was at 2 [m] away from the camera. A depth template is a binary template; the foreground and the background value are adjusted according to the status of tracks and input data.

Figure 1. Depth templates.



(a) Input image (b) Depth image

Figure 2. Person detection result.

For a person being tracked, his/her scene position is predicted using the Kalman filter. We thus set the foreground depth of the template to the predicted depth of the head of the person. Then we calculate the dissimilarity between a depth template and the depth image using an SSD (sum of squared distances) criterion.

To detect a person in various orientation, we use the three templates simultaneously and take the one with the smallest dissimilarity as a matching result. An example of detection using the depth templates is shown in Figure 2. We set a detection volume to search in the scene; its height range is $0.5 \sim 2.0$ [m] and the range of the depth from the camera is $0.5 \sim 5.5$ [m].

2.2 SVM-based false rejection

A simple template-based detection is effective in reducing the computational cost but at the same time may produce many false detections for objects with similar silhouette to person. To cope with this, we use an SVM-based person verifier when person candidates are detected.

We collected many person candidate images in various orientation detected by the depth templates. We used 356 positive and 147 negative images for training. A person candidate region in the image is resized to 40×40 [pixels] to generate a 1600-dimensional intensity vector. HOG features [1] for that region are summarized into a 2916-dimensional vector. These two vectors are concatenated to generate a 4516-dimensional feature vector, which is used for training and classification.

2.3 EKF-based tracking

We adopt the Extended Kalman Filter (EKF) for robust data association and occlusion handling. The

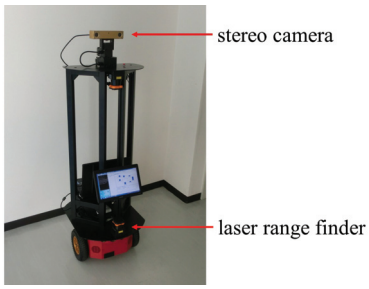


Figure 3. A mobile robot with a laser range finder and a stereo camera.

state vector includes the position and the velocity in the horizontal axes (X and Y) and the height (Z) of a person. The vector is represented in the robot local coordinates and a coordinate transformation is performed from the previous to the current robot's pose every time in the prediction step, using the robot's odometry information.

Color information of the clothing is used for identifying the target person to follow.

3 Person following robot

3.1 Configuration of the system [8]

We deal with two kinds of objects in the environment: persons detected by stereo vision and static obstacles detected by a laser range finder (LRF). The system consists of the following three modules:

1) Person detection and tracking module detects persons using stereo and tracks using Kalman filtering to cope with occasional occlusions among people. Details of the processing are described in Sec. 2.

2) Local map generation module constructs and maintains an occupancy grid map, centered at the current robot position, using the data from the LRF. It performs a cell-wise Bayesian update of occupancy probabilities assuming that the odometry error can be ignored for a relatively short robot movement.

3) Motion planning module calculates a safe robot motion whitch follows a specified target person and avoids others, using a randomized kinodynamic motion planner.

Note that we use the term "tracking" in the image processing, and use "following" in the robot's movement.

3.2 Hardware configuration

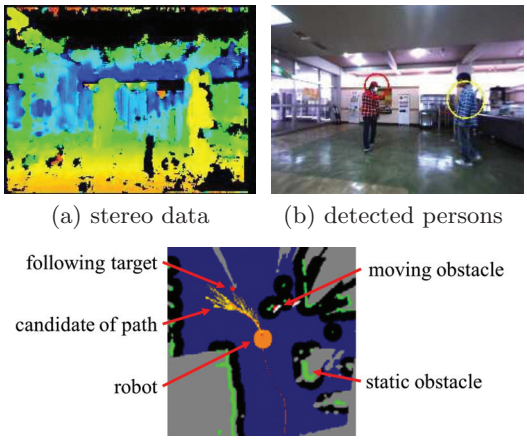
Figure 3 shows our mobile robot system which is composed of a robot base (Pioneer 3-DX by Mobile Robots), a stereo camera (Bumblebee2 by Point Grey Research), a laser range finder (UTM-30LX by Hokuyo), and a Note PC (Core i7, 2.70GHz).

3.3 Example of recognition and path planning

Figure 4 shows an example of the scene recognition and the planning result. From the stereo data (see Fig. 4(a)), the robot detected two persons, the target on the left and the other on the right (see Fig. 4(b)). Figure 4(c) shows the result of scene recognition and motion planning.

4 Experimental Result

We performed experiments of specific person following at the exhibitions in which many ordinary people exist. Table 1 shows the experimental conditions. We did experiments not only in indoor but also in outdoor.



(a) stereo data (b) detected persons

(c) recognition result and path planning

Figure 4. An example of scene recognition and motion planning.

Our system was able to work well even in outdoor (see Fig. 5(c)). In the case that most of businessmen or researchers wear the suit, it is easy to identify the target person who wears the clothing of a distinctive color. On the other hand, it may sometimes be difficult to distinguish the target from ordinary people who wear various clothing. In addition, people tended to avoid the robot in open spaces, while people often passed between the robot and the target person in narrow passages. We think that the environment of (d) the exhibition "MONOZUKURIHAKU" is similar to usual shopping malls, therefore we analyze the results.

Figures 5 and 6 show snapshots of the person following experiments. Figure 6(a) is an example of person tracking result. Note that each circle shows a tracked person, and the red circle indicates the target person.

4.1 Result of person following experiment

Table 2 shows the result of person following experiment in the exhibition "MONOZUKURIHAKU". The processing speed was about 7 [fps]. The total number of frames was 82454 (about 200 minutes). The target person was successfully tracked with 97.8% of frames. The number of failure in image processing was 112. There were the following three cases in failure:

a) Failures by occlusion

The number of occlusions was 183, in 134 cases of which the target person was tracked correctly as shown in Figure 7(a). In the remaining 49 cases, another person was tracked as shown in Figure 7(b). The target person, however, was detected and tracked when the person appeared again in the image.

Note that this case includes intentional occlusions in which a person blocked the robot's course (Fig. 8(a)) or covered the robot's camera (Fig. 8(b)). In these cases, the target person was occluded more for a long time compared with simple intersection.

b) Failures by wrong depth information

There were ten cases where a static object was tracked (Fig. 7(c)) because of wrong depth information caused by failure of stereo processing.

c) Failures by misidentification

The number of failures in the target identification was 53. Figure 7(d) shows the failure of target identi-

Table 1. Experimental conditions of specific person following in real environment.

event	date	main participants	location	space
(a) the exhibition "Business Link"	Jan. 19, 2012	businessman	indoor	open
(b) the conference "ROBOMEC2012"	May 28∼29, 2012	researcher	indoor	passage
(c) open campus	Aug. 25, 2012	ordinary family	outdoor	open
(d) the exhibition "MONOZUKURIHAKU"	Nov. 30∼Dec. 1, 2012	ordinary family	indoor	passage

Table 2. Result of person following experiment in the exhibition "MONOZUKURIHAKU".

result \ response	tracking recovered automatically	target returned to robot's front	emergency stop	total
successfully tracked	—	—	(3 times)	80669 frames (97.8%) / —
failure by occlusion	609 frames / 47 times	78 frames / 2 times	0 times	687 frames (0.8%) / 49 times
failure by wrong depth	59 frames / 9 times	22 frames / 1 times	0 times	81 frames (0.1%) / 10 times
failure by misidentification	278 frames / 30 times	739 frames / 23 times	0 times	1017 frames (1.2%) / 53 times
total	946 frames / 86 times	839 frames / 26 times	(3 times)	82454 frames (100%) / 112 times

(a) the exhibition "Business Link"

(b) the conference "ROBOMEC2012"

(c) open campus

Figure 5. Snapshots of person following experiments in real environment.

fication because another person with a similar color of clothing exists near the target person.

For robust person identification, it is necessary to use various cues together. We are developing a SIFT feature-based identification method [9] which uses not only the color but also the texture of clothing. However, it is necessary to additionally use other features because it cannot be used for a clothing with no texture.

4.2 Response to the tracking failure

We classified 112 failures into the following three cases according to the target person's response:

1) The tracking was recovered automatically within about 1.5 [s] (an average of 11 frames) in 86 cases. The robot continued to follow the target person without his noticing the failure.

2) The target person had to return to the robot's front in 26 cases, because the robot turned to the mistaken direction. Person following continued without a restart of the system.

3) Emergency stop caused by a failure in image processing was not performed. When a person approaches from the front, the robot goes back. We stopped the robot only when there was danger of a collision because the robot cannot perceive objects in the back.

4.3 Following a ordinary person

Figure 9 shows examples of the experiment in which the robot followed ordinary people. A child was also successfully tracked.

5 Conclusions

This paper presented an experimental evaluation of specific person following by a vision-based mobile robot in the situation where many ordinary people exist. The target person was successfully tracked with 97.8% of frames, and the robot continued to follow without a restart of the system. However, misidentification of the target sometimes occurred. We will improve the person identification ability which additionally uses other personal features.

Acknowledgment

A part of this research is supported by JSPS KAKENHI 23700203 and NEDO Intelligent RT Software Project.

References

[1] N. Dalal and B. Triggs: "Histograms of oriented gradients for human detection," *CVPR2005*, pp. 886–893, 2005.

[2] L. Spinello and K.O. Arras: "People Detection in RGB-D Data," *IROS2011*, pp. 3838–3843, 2011.

[3] M. Munaro, et al.: "Tracking people within groups with RGB-D data," *IROS2012*, pp. 2101–2107, 2012.

[4] J. Salas and C. Tomasi: "People Detection Using Color and Depth Images," *MCPR2011*, pp. 127–135, 2011.

[5] A. Ess, et al.: "Object Detection and Tracking for Autonomous Navigation in Dynamic Environments," *IJRR*, vol. 29, no. 14, pp. 1707–1725, 2010.

[6] S. Frintrop, et al.: "A Component-based Approach to Visual Person Tracking from a Mobile Platform," *Int. J. Social Robotics*, vol. 2, no. 1, pp. 53–62, 2010.

[7] J. Satake and J. Miura: "Robust Stereo-Based Person Detection and Tracking for a Person Following Robot," *ICRA-2009 Workshop on People Detection and Tracking*, 2009.

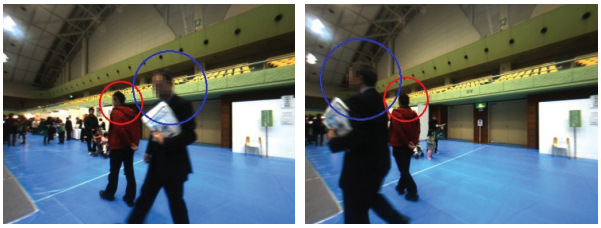
[8] J. Miura, et al.: "Development of a Person Following Robot and Its Experimental Evaluation," *IAS-11*, pp. 89–98, 2010.

[9] J. Satake, et al.: "A SIFT-Based Person Identification using a Distance-Dependent Appearance Model for a Person Following Robot," *ROBIO2012*, pp. 962–967, 2012.

(a) tracking result (robot view)

(b) snapshots of person following (outside view)

Figure 6. An example of person following in the exhibition “MONOZUKURIHAKU”.



robot view robot view

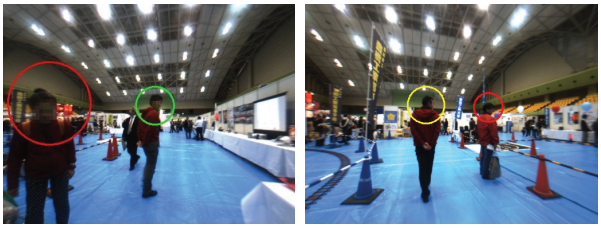
(a) successfully tracked

robot view robot view

(b) failure by occlusion

robot view robot view

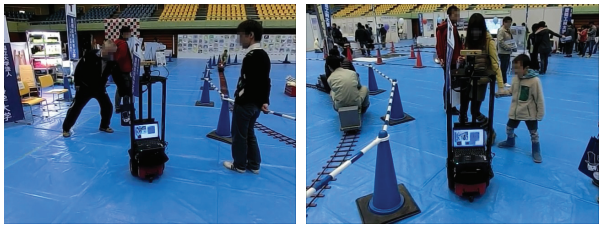
(c) failure by wrong depth information



robot view robot view

(d) failure by misidentification

Figure 7. Examples of image processing result.



outside view outside view

(a) persons blocked the robot’s course

robot view outside view

(b) a person covered the robot’s camera

Figure 8. Intentional disturbance.

robot view outside view

(a) follow a woman

robot view outside view

(b) follow a child

Figure 9. The robot followed ordinary people.